SEQUENTIAL RESOURCE ALLOCATION UNDER MULTI-ARMED BANDIT MODEL WITH ONLINE CLUSTERING AS SIDE INFORMATION

A. Dzhoha[1] **and I. Rozora**[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: andrew.djoga@gmail.com

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: irozora@knu.ua

Abstract

We consider the sequential resource allocation problem under the multi-armed bandit model in the non-stationary stochastic environment. The stochastic multi-armed bandit problem is a classic example of the exploration-exploitation dilemma which is originally presented by Thompson (1933) in the context of clinical trials and later formalized by Robbins (1952). It's a sequential problem defined by a set of actions where at each step, an action is selected, and then a stochastic environment reveals a reward. The goal is to maximize the total reward obtained in a sequence after all steps. Motivated by many real applications, where information can naturally be grouped, we consider a variation of the contextual multi-armed bandit (Bubeck & Cesa-Bianchi 2012) with online clustering representing side information. We assume a stochastic environment, in which the reward of each action conditioned on a cluster follows a Bernoulli distribution with unknown parameters. Additionally, we assume that the nature of the problem changes over time and the clusters drift incrementally making the reward process non-stationary. In this setting, we propose a new algorithm based on a two-stage approach. The first stage is a sequential modification of the traditional k-means clustering algorithm (Duda et al. 2001), in which the algorithm deals with the continuous data stream and acts on a subset of data rather than in a single batch. In the second stage, we incorporate the current information about clusters into the Thompson Sampling algorithm, which is one of the stochastic bandit policies. We introduce a discounting mechanism to track changes in the underlying reward and account for a potential cluster misclassification. We provide the regret analysis of edge cases with supporting numerical experiments for this algorithm.

Keywords: multi-armed bandit problem, non-stationary stochastic environment, online clustering.

References

Bubeck S., Cesa-Bianchi N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122.

Duda R. O., Hart P. E. and Stork D. G. (2001) *Pattern Classification.* John Wiley & Sons.

Robbins, H. (1952) Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, **58**(5), 527–535.

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.